International Review on Computers and Software (I.RE.CO.S.), Vol. 4, n. 1

Arabic Statistical N-gram Models

K. Meftouh[1], K. Smaili[2], M. T. Laskri[1]

Abstract – *In this work we propose to investigate statistical language models for Arabic. Several experiments using different smoothing techniques have been carried out on a small corpus extracted from a daily newspaper. The sparseness data conducts us to investigate other solutions without increasing the size of the corpus. A word segmentation has been operated in order to increase the statistical viability of the corpus. This leads to a better performance in terms of normalized perplexity*

***Keywords**: Statistical language model, Arabic, N-gram models, word-based n-gram models, Morpheme-based n-gram models, Perplexity.*

Nomenclature

K	The number of words constituting a sequence.
W_i	The word to be predicted
P(e)	The probability measure which indicates the likelihood of the event e.
$P(w_i/h)$	The probability assigned to the word w_i in its context h.
UNK	Any word not in the vocabulary is replaced in the corpora by an abstract entity noted UNK which means Unknown word.
E	Entropy: A measure used to evaluate the quality of a statistical model.
PP	Perplexity: It can be viewed as an average size of words which can follow a phrase.
PP_n	The normalized perplexity.

I. Introduction

Statistical techniques have been widely used in automatic speech recognition and machine translation over the last two decades [1]. Most of the success, therefore, has been witnessed in the so called “resource rich languages” for instance English and French. More recently there has been an increasing interest in languages such as Arabic.

Arabic has a rich morphology characterized by a high degree of affixation and interspersed vowel patterns and roots in word stems, as shown in section 2. As in other morphologically rich languages, the large number of possible word forms entails problems for robust language model estimation.

A statistical language model is used to build up sequence of words, classes or phrases which are linguistically valid without any use of external knowledge. A list of probabilities is estimated from a large corpus to indicate the likelihood of linguistic events. An event is any potential succession of words. The common model used in the literature is the well known n-grams. A word is estimated in accordance to the $(n-1)$ previous words. To be efficient this model needs a huge amount of data to train all the needed parameters. Due to the relative recent interest for Arabic applications, the necessary resources for this language are not as important as what we have for the Indo-European languages. In the present work, we investigate several classical statistical language models in order to study their pertinence for Arabic language. Sparseness data conducts us to test several smoothing techniques in order to find out the best model. In the following section, we will give an overview of Arabic language (section 2). We pursue by a description of the n-gram models (section 3), then the used corpora (section 4). In sections 5 and 6 we present respectively the results obtained with word-based N-gram models and morpheme models. Finally we conclude.

II. An overview of Arabic

Arabic, one of the six official languages of the United Nations, is the mother tongue of 300 millions people [2]. Unlike Latin-based alphabets, the orientation of writing in Arabic is from right to left. The Arabic alphabet consists of 28 letters and can be extended to ninety by additional shapes, marks and vowels. Each letter can appear in up to four different shapes, depending on whether it occurs at the beginning, in the middle, at the end of a word, or alone. Table 1 shows an example of the letter < ف /”f’ > in its various forms. Letters are mostly connected and there is no capitalization.

Manuscript received January 2007

TABLE1
THE LETTER < ف/”f’ > IN ITS VARIOUS FORMS

Isolated	Beginning	Middle	End
ف	فـ	ـفـ	ـف

Arabic is a Semitic language. The grammatical system of Arabic language is based on a root-and-pattern structure and considered as a root-based language with not more than 10000 roots and 900 patterns [3]. The root is the bare verb form. It is commonly three or four letters and rarely five. Pattern can be thought of as template adhering to well-known rules.

Arabic words are divided into nouns, verbs and particles. Nouns and verbs are derived from roots by applying templates to the roots to generate stems and then introducing prefixes and suffixes [4]. Table 2 lists some templates (patterns) to generate stems from roots. The examples given below are based on the root درس/ < drs > (Refer to table 1 in the Appendix for the mapping between the Arabic letters and their Latin representations).

TABLE2
SOME TEMPLATES TO GENERATE STEMS FROM THE ROOT درس/ < drs >. C INDICATE A CONSONANT A a VOWEL.

Template	Stem
فعل CCC	درس < drs >/ Study
فاعل CACC	دارس < dArs >/ Student
مفعول mCCwC	مدروس < mdrws >/ Studied

Many instances of prefixes and suffixes correspond to entire words in other languages. In table 3, we present the different components of a single word وكررتها which corresponds to the phrase "and she repeats it".

TABLE3
AN EXAMPLE OF AN ARABIC WORD

French	Arabic	English
et	و	And
répéter	كرر	Repeat
elle	ت	She
la	ها	It

Arabic contains three genders (much like English): masculine, feminine and neuter. It differs from Indo-European languages in that it contains three numbers instead of the common two numbers (singular and plural). The third one is the dual that is used for describing the action of two people.

III. N-gram Models

The goal of a language model is to determine the probability $P(w_1 \ldots w_k)$ of a word sequence $w_1 \ldots w_k$. This probability is estimated as follows:

$$P(w_1,\ldots,w_k) = \prod_{i=1}^{k} P(w_i / w_1,\ldots,w_{i-1}) \qquad (1)$$

The most widely-used language models are n-gram models [5]. In n-gram language models, we condition the probability of a word on the identity of the last $(n-1)$ words.

$$P(w_1,\ldots,w_k) = \prod_{i=1}^{k} P(w_i / w_{i-n+1},\ldots,w_{i-1}) \qquad (2)$$

The choice of n is based on a trade-off between detail and reliability, and will be dependent on the available quantity of training data [5]. Because of the sparseness data, in statistical language models, parameters have to be smoothed. The objective is to fine-tune probabilities to overcome the problem of missing data. Several methods exist in the literature.

IV. Data Description

The experiments reported, in this section, are conducted on corpora extracted from Al-khabar (an Algerian Daily newspaper). Al-Khabar is written on modern standard Arabic, the one used by all the official media in Arabic world. One of the specificity of Arabic language is that a text can be read without using any vowel. That is why articles in newspapers are unvocalized. The corpora we use contain 80K words for training and 5K words for test. Figure 1 shows a sample of the training corpus.

سوق اهراس
انجراف التربة يستهلك الملايير دون
جدوى عجزت المصالح التقنية لولاية
سوق اهراس عن مجابهة مشكل انجراف
التربة الذي يلحق أضرارا كبيرة
بشتى القطاعات وتنجر عنه خسائر
مالية هامة تكون على حساب متطلبات
تنموية أخرى وهو ما جعل الولاية
من أكثر الولايات تخلفا وأكثرها
تأخرا في المرافق الإجتماعية
الضرورية

Fig.1. A sample of the training corpus

For the both following experiments the language models have been smoothed by three techniques: Good-Turing [6], Witten-Bell [7] and linear [8].

V. Word-Based N-gram Models

The baseline model is calculated with a vocabulary of the most frequent 2000 words. The UNK may distort the interpretation of results because of its occurrence, so it can act in favor of a better language model within the meaning of perplexity if the vocabulary has a weak cover. Table 4 and table 5 show the performance in terms of test perplexity without and including UNK respectively. The rate of unknown words is 30.19 %.

TABLE4
PERPLEXITY AND ENTROPY PERFORMANCE WITHOUT UNK.

n	Good-Turing		Witten-Bell		Linear	
	PP	E	PP	E	PP	E
2	289.10	8.18	267.86	8.07	309.29	8.27
3	292.36	8.19	278.87	8.12	321.50	8.33
4	307.51	8.26	311.97	8.29	335.14	8.39

TABLE5
PERPLEXITY AND ENTROPY PERFORMANCE INCLUDING UNK.

n	Good-Turing		Witten-Bell		Linear	
	PP	E	PP	E	PP	E
2	76.66	6.26	76.03	6.25	82.92	6.37
3	81.55	6.35	81.18	6.35	92.09	6.52
4	88.07	6.46	89.25	6.48	97.67	6.61

Indeed more the corpus contains UNK, more the probability of this fictive word is large, which leads to a less perplexity. It is thus desirable to always calculate perplexity without UNK.

Note also that the values of perplexity are high and increase according to the order of the model n. This is due to the weak size of the training corpus. To take into account the sparseness data issue, we propose to split words into morphemes. This operation leads to increase the frequency of basic units and consequently to reduce the percentage of unknown words.

VI. Morpheme-Based N-gram Models

Languages with rich morphology generate so many representations from the same root. Often, this makes them highly flexional and consequently the perplexity could be important [9]. An Arabic word consists of a sequence of morphemes respecting the following pattern prefix*-stem-suffix* (* denotes zero or more occurrences of a morpheme). We define an n-morpheme model as an n-gram of morphemes. In this case the corpus is rewritten in terms of morphemes as in the example of figure 2.

سوق_اهراس
انجراف ال تربة يستهلك ال ملايير
دون جدوى عجزت ال مصالح ال تقنية
لـ ولاية سوق_اهراس عن مجابهة مشكل
انجراف ال تربة ال ذي يلحق أضرارا
كبيرة بـ شتى ال قطاعات و
تنجر عنه خسائر مالية هامة تكون على
حساب متطلبات تنموية أخرى و
هو ما جعل ال ولاية من أكثر ال
ولايات تخلفا و أكثرها تأخرا في ال
مرافق ال إجتماعية ال ضرورية

Fig.2. A sample of Arabic morphemes corpus

The quality of a language model is estimated by the test perplexity PP :

$$PP = 2^{-\frac{1}{N}\log_2(P(W))} \tag{3}$$

With N is the size of the test corpus.
When we proceed to a decomposition of words into *prefix*-stem-suffix**, we modify the number of items constituting the original corpus W. To make the comparison of the two models relevant, the perplexity has to be normalized [10] as follows:

$$PP_n = 2^{\frac{N_1}{N_2}\log_2(PP)} \tag{4}$$

where N_1, N_2 correspond respectively to the size of the original corpus and the rewritten one.

The prefixes which are used for the segmentation are the common used in Arabic language.

TABLE6
PREFIXES AND THEIR MEANINGS

Prefixes			
“w” و	and	“l”لـ	to
“k”كـ	like	“b”بـ	with
“f”فـ	then	“Al” الـ	the

To make the corpus statistically reliable and to fit the reality of the Arabic language, some words have been gathered. That is why for instance, we concatenate the town's name composed by two or more words [11]. See Table 7 for an example. This operation is handled by using a predefined list of composed words. Work is under progress to find out automatically sequence of Arabic words.

TABLE7
AN EXAMPLE OF COMPOSED TOWN'S NAME.

سوق أهراس Is rewritten: سوق_أهراس

The transformation of the initial corpora leads to respectively a training and a test corpus of 110K and 6,9K tokens. Table 8 and table 9 show the values of the unnormalized perplexity without and including UNK.

TABLE8
PERPLEXITY AND ENTROPY PERFORMANCE WITHOUT UNK.

n	Good-Turing		Witten-Bell		Linear	
	PP	E	PP	E	PP	E
2	87.89	6.46	86.44	6.43	94.25	6.56
3	68.42	6.10	65.44	6.03	75.50	6.24
4	69.82	6.13	66.70	6.06	76.08	6.25

TABLE9
PERPLEXITY AND ENTROPY PERFORMANCE INCLUDING UNK.

n	Good-Turing		Witten-Bell		Linear	
	PP	E	PP	E	PP	E
2	57.63	5.85	57.66	5.85	61.91	5.95
3	47.27	5.56	46.17	5.53	52.81	5.72
4	48.72	5.61	47.11	5.56	53.96	5.75

We remark that 3-gram and 4-gram models lead to better results than bigram. This is due to the fact that this segmentation makes the corpus statistically viable. Indeed, the decomposition decreases the variety of bigrams and increase the frequency of tree and four grams. In order to compare the perplexity to that obtained with the original language models, we compute the normalized perplexity. Table 10 lists these values.

TABLE10
NORMALIZED PERPLEXITY'S VALUES.

n	Good-turing	Witten-Bell	Linear
2	173.52	170.18	188.05
3	130.05	123.55	145.61
4	133.06	126.23	146.93

These results show an improvement of 55.7% in terms of 3-gram perplexity using Witten-Bell smoothing technique. We can state that for small corpus, the segmentation of words improve the language model and this, whatever the used technique of smoothing.

VII. Conclusion

In this work we used n-grams to model Arabic language; several experiments have been carried out on a small corpus extracted from a daily newspaper. The sparseness data conducts us to investigate other solutions without increasing the size of the corpus. We think that even with a large corpus, segmentation is necessary. In fact, a lot of words in Arabic are constructed from patterns which are used as generative rules. Each pattern indicates not only how to construct a word but gives the syntactic role of the generated word.
Several experiments and developments are under work, the objective is to obtain a very robust statistical Arabic language model. Among them, we can mention

- A tool capable of segmenting a word into *prefix*-stem-suffix*.*
- An evaluation of a complete morpheme-based n-gram model.
- Arabic n-class model.
- A dynamic Bayesian Network formalism for Arabic statistical modeling.

Appendix

Table 1
LETTER MAPPINGS

A	d	m	r	S	w
ا	د	م	ر	س	و

References

[1] Kim, W., Khudanpur, S. 2003. Cross-Lingual lexical triggers in statistical language modelling. *Theoretical Issues In Natural Language Processing archive.* Proceedings of the 2003 conference on Empirical methods in natural language processing, Vol. 10

[2] Egyptian Demographic center. (2000). http://www.frcu.eun.eg/www/homepage/cdc/cdc.htm

[3] Hayder K. Al Ameed, Shaikha O. Al Ketbi and al. (2005). Arabic light stemmer: A new enhanced approach. *Proc. of IIT'05 (the Second International Conference on Innovations in Information Technology)*

[4] Darwish K. (2002). Building a shallow Arabic morphological analyser in one day. *Proc. of the ACL workshop on computational approaches to Semitic languages.*

[5] Stanley F.Chen, Goodman J. (1998). An empirical study of smoothing techniques for language modelling. *Technical report TR-10-98, Computer science group*, Harvard University, Cambridge, Massachusetts.

[6] Katz S.M. (1987). Estimation of probabilities from sparse data for the language model component of a speech recognizer.

IEEE Transactions on Acoustics, Speech and Signal processing, 35(3): 400-401.

[7] Witten I.T. and Bell T.C. (1991). The Zero-frequency problem: Estimating the probabilities of novel events in adaptive text compression. *IEEE Transactions on Information Theory*, 37(4):1085-1094.

[8] Ney H., Essen U. and Kneser R. (1994). On structuring probabilistic dependencies in stochastic language modeling. *Computer Speech and Language*, 8(1):1-38.

[9] Kirchoff K. et al. (2002). Novel approches to Arabic speech recognition. *Technical report. Final report from the 2002 Johns-Hopkins summer workshop*. John-Hopkins university.

[10] Gauvain J.L., Lamel L., Adda G, and Matrouf D. (1996). The LIMSI 1995 Hub system. *Proc. ARPA Spoken Language Technology Workshop-96*.

[11] Zitouni I., Smaili K. and J.P. Haton (2003). Statistical Language Modeling Based on Variable-Length Sequences, *Computer Speech and Language*.

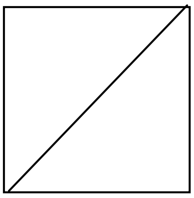
Authors' information

[1]Badji Mokhtar University, Computer Science Department, BP 12

23000 Annaba, Algeria

[2]INRIA, LORIA, Parole team, BP 101 54602 Villers Les Nancy,

France



Karima Meftouh was born at Annaba (Algeria) in 1968. She studied at Annaba University where she received the degree of computer science engineer in 1992, the Master degree in 2000. Since 2001, she is a research teacher at the computer science department of Annaba University and a member of the research group in artificial intelligence within the LRI laboratory. Currently Mrs. Meftouh prepares her Doctorate in the field of the computational linguistics. She is interested particularly in the Arabic statistical language modeling. Her research was the several publication object in various conferences: CITALA' 2007, SIIE' 2008, JADT' 2008, ICAART' 2009.

Kamel Smaili Professor at university of Nancy since 2002 obtained a PHD from the same university on 1991. He is a member of LORIA-France lab. He is a leader of a research group working on statistical language modeling and speech-to-speech translation. He defended an HDR (Habilitation à diriger la recherche) on 2001. His research interest since 20 years concerns statistical language modeling for speech recognition and since 2000 he oriented his research to speech-to-speech translation. He proposed several original ideas: retrieving phrases based on class-phrases, purging statistical language models from impossible events, Cache-features language model, multilingual triggers, . . .He participated to several European and French projects concerning speech recognition: COCOS, MULTIWORKS, COST, MIAMM, IVOMOB (RNRT project). He advised more than 9 PHD students and participated to 20 PHD committees through the world. He took part to several program committees: Eurospeech, ICSLP, ICASSP, SIIE, TAIMA, TAL, Computer speech and language, Speech communication, . . .He published his research in more than 55 international conferences and journals and in more than 20 francophone conferences and journals.

Mohamed Tayeb Laskri was born at Annaba (Algeria) in 1958. He obtained a 3rd cycle doctorate on computer science (France, 1987) He obtained his PhD degree from Annaba University (Algeria, 1995). Pr. Laskri is head of the research group in artificial intelligence within LRI laboratory. His current research takes the reasoning in artificial intelligence like field of application privileged in particular in image processing, multi-agents systems, engineering of the Human-machine interfaces and automatic natural language processing.